# Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes

Rausell Antonio[1,2,§,*], Luo Yufei[1,2,§], Lopez Marie[3], Seeleuthner Yoann[2,4], Rapaport Franck[5], Favier Antoine[1,2], Stenson Peter D.[6], Cooper David N.[6], Patin Etienne[3], Casanova Jean-Laurent[2,4,5,7,8], Quintana-Murci Lluis[3,9] , Abel Laurent[2,4,5,‡,*]

1. Clinical Bioinformatics Laboratory, INSERM UMR1163, Necker Hospital for Sick Children, 75015 Paris, France, EU

2. Paris University, Imagine Institute, 75015 Paris, France, EU

3. Human Evolutionary Genetics Unit, Institut Pasteur, UMR2000, CNRS, Paris 75015, France, EU

4. Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM UMR1163, Necker Hospital for Sick Children, 75015 Paris, France, EU

5. St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY, USA

6. Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff CF14 4XN, UK, EU.

7. Howard Hughes Medical Institute, New York, NY, USA

8. Pediatric Hematology and Immunology Unit, Necker Hospital for Sick Children, 75015 Paris, France, EU.

9. Human Genomics and Evolution, Collège de France, Paris 75005, France, EU

§ Joint first authors, equal contributions

* Correspondence to

antonio.rausell@inserm.fr

casanova@mail.rockefeller.edu

laurent.abel@inserm.fr

Keywords: Redundancy, pseudogenization, loss of function, positive selection, negative selection

**Abstract**. Humans homozygous or hemizygous for variants predicted to cause a loss of function (LoF) of the corresponding protein do not necessarily present with overt clinical phenotypes. We report here 190 autosomal genes with 207 predicted LoF variants, for which the frequency of homozygous individuals exceeds 1% in at least one human population from five major ancestry groups. No such genes were identified on the X and Y chromosomes. Manual curation revealed that 28 variants (15%) had been misannotated as LoF. Of the 179 remaining variants in 166 genes, only 11 alleles in 11 genes had previously been confirmed experimentally to be LoF. The set of 166 dispensable genes was enriched in olfactory receptor genes (41 genes). The 41 dispensable olfactory receptor genes displayed a relaxation of selective constraints similar to that observed for other olfactory receptor genes. The 125 dispensable non-olfactory receptor genes also displayed a relaxation of selective constraints consistent with greater redundancy. Sixty two of these 125 genes were found to be dispensable in at least three human populations, suggesting possible evolution toward pseudogenes. Of the 179 LoF variants, 68 could be tested for two neutrality statistics, and eight displayed robust signals of positive selection. These variants included a known *FUT2* variant that confers resistance to intestinal viruses, and an *APOL3* variant involved in resistance to parasitic infections. Overall, the identification of 166 genes for which a sizeable proportion of humans are homozygous for predicted LoF alleles reveals both redundancies and advantages of such deficiencies for human survival.

**Significance statement.** Human genes homozygous for apparent loss of function (LoF) variants are increasingly reported in a sizeable proportion of individuals without overt clinical phenotypes. Here, we found 166 genes with 179 predicted LoF variants for which the frequency of homozygous individuals exceeds 1% in at least one of the populations present in databases ExAC and gnomAD. These putatively dispensable genes showed relaxation of selective constraints suggesting that a considerable proportion of these genes may be undergoing pseudogenization. Eight of these LoF variants displayed robust signals of positive selection, including two variants in genes involved in resistance to infectious diseases. The identification of dispensable genes will facilitate the identification of functions that are now redundant, or possibly even advantageous, for human survival.

**/body**

## Introduction

The human genome displays considerable DNA sequence diversity at the population level. One of its most intriguing features is the homozygosity or hemizygosity for variants of protein-coding genes predicted to be loss-of-function (LoF) found at various frequencies in different human populations (1–3). An unknown proportion of these reported variants are not actually LoF, instead being hypomorphic or isomorphic, because of a re-initiation of translation, readthrough, or a redundant tail, resulting in lower, normal, or even higher than normal levels of protein function. Indeed, a *bona fide* nonsense allele, predicted to be LoF, can actually be gain-of-function (hypermorphic), as illustrated by IκBα mutations (4). Moreover, the LoF may apply selectively to one isoform or a subset of isoforms of a given gene, but not others (e.g. if the exon carrying the premature stop is spliced out for a specific set of alternative transcripts) (5). Finally, there are at least 400 discernible cell types in the human body (6), and the mutant transcript may be expressed in only a limited number of tissues. Conversely, there are also mutations not predicted to be LoF, such as in-frame insertions-deletions (indels), missense variants, splice-region variants affecting the last nucleotides (nt) of exons and even synonymous or deep intronic mutations, that may actually be LoF but cannot be systematically identified as such *in silico*.

Many predicted LoF variants have nevertheless been confirmed experimentally, typically by demonstration of their association with a clinical phenotype. Of the 229,161 variants reported in HGMD (7), as many as 99,027 predicted LoF alleles in 5,186 genes have been found to be disease-causing in association and/or functional studies. For example, for the subset of 253 genes implicated in recessive forms of primary immunodeficiencies (8), 12,951 LoF variants are reported in HGMD. Conversely, a substantial proportion of genes harboring biallelic null variants have no discernible associated pathological phenotype, and several large-scale sequencing surveys in adults from the general population have reported human genes apparently tolerant to homozygous LoF variants (9–14). Four studies in large bottlenecked or consanguineous populations detected between 781 and 1,317 genes homozygous for mutations predicted to be LoF (10, 11, 13, 14). These studies focused principally on low-frequency variants (minor allele frequency, MAF<1%-5%), and associations with some traits, benign or disease-related, were found for a few rare homozygous LoF variants. Two studies provided a more comprehensive perspective of the allele frequency spectrum of LoF variants.

A first systematic survey of LoF variants, mostly in the heterozygous state, was performed with whole-genome sequencing data from 185 individuals of the 1000 Genomes Project; it identified 253 genes with homozygous LoF variants (9). In a larger study of more than 60,000 whole exomes, the ExAC project identified 1,775 genes with at least one homozygous LoF variant, with a mean of 35 homozygous potential LoF variants per individual (12).

These studies clearly confirmed the presence of genes with homozygous LoF variants in apparently healthy humans, but no specific study of such variants present in the homozygous state in a sizeable proportion (>1%) of large populations has yet been performed. In principle, these variants may be neutral (indicating gene redundancy) (15), or may even confer a selective advantage (the so-called “less is more” hypothesis) (3, 16). Indeed, a few cases of common beneficial LoF variants have been documented, including some involved in host defense against life-threatening microbes (17, 18). Homozygosity for LoF variants of *DARC (*now *ACKR1), CCR5*, and *FUT2* confer resistance to *Plasmodium vivax (19, 20)*, HIV (21–23), and norovirus (24, 25), respectively. We hypothesized that the study of common homozygous LoF variants might facilitate the identification of the set of dispensable protein-coding genes in humans and reveal underlying evolutionary trends. Unlike rare LoF variants, *bona fide* common homozygous LoF variants are predicted to be enriched in neutral alleles (1–3). They also provide indications concerning genes undergoing inactivation under positive selection (17, 18). The availability of large public databases, such as ExAC (http://exac.broadinstitute.org/, (12)) and its extended version gnomAD (http://gnomad.broadinstitute.org/), which includes data from more than 120,000 individuals, is making it possible to search for such variants with much greater power, across multiple populations.

## Results

### Definition of the set of dispensable protein-coding genes in humans

We used two large exome sequence databases: the Exome Aggregation Consortium database (ExAC; (12)) and the Genome Aggregation Database (GnomAD; **Methods**), which have collected 60,706 and 123,136 exome sequences, respectively. For this study, we focused on the 20,232 protein-coding genes and we excluded the 13,921 pseudogenes (**Methods**; (26)). We defined protein-coding genes as dispensable if they carry variants that: (i) are computationally predicted to be loss-of-function

(LoF) with a high degree of confidence, including early stop-gains, indel frameshifts, and essential splice site-disrupting variants (i.e. involving a change in the 2-nt region at the 5' or 3' end of an intron; **Methods**); and (ii) have a frequency of homozygous individuals (hemizygous for genes on the X chromosome in males) exceeding 1% in at least one of the five population groups considered in these public databases (i.e. Africans, including African Americans, East Asians, South Asians, Europeans, and Admixed Latino Americans). As we focused on exome data, only small indels (<50 bp) were considered in this analysis. Common quality filters for calls and a minimum call rate of 80% were applied to each reference database (ExAC and GnomAD; **Methods**). Only LoF variants affecting the principal isoform (27) were retained (**Methods**). The application of these filters led to the detection of 208 (ExAC) and 228 (GnomAD) genes, 190 of which were common to the two databases, and are referred to hereafter as the *set* of dispensable genes (**Table 1** and **Supplementary Table 1**). No genes on the X or Y chromosomes fulfilled these criteria. Relaxing the thresholds on the SNP and INDEL call quality filters (variant quality score recalibration (VQSR) score; **Methods**), the variant call rate or the non-restriction to the principal isoform did not substantially increase the number of putatively dispensable genes (**Supplementary Figures 1** and **2)**. The frequency of homozygous individuals at which a gene is defined as dispensable appeared to be the criterion with the largest impact on the number of dispensable genes identified, thereby justifying the use of the stringent threshold (>1% homozygotes in at least one specific population) described above (**Supplementary Figures 1** and **2)**.

**Manual curation of the LoF variants**

We then investigated the potential molecular consequences of the 207 putative LoF variants retained, collectively associated with the 190 genes, by manually curating the annotation of these LoF variants. An examination of the sequencing reads in GnomAD showed that 18 variants — 10 single nucleotide variants (SNVs) and eight indels initially annotated as stop-gains and frameshift variants, respectively — were components of haplotypes with consequences other than the initial LoF annotation (**Supplementary Table 1**). Thus, for each of the 10 SNVs, a haplotype encompassing a contiguous second variant led to the creation of a missense rather than a stop variant allele. For each of the eight frameshift variations, a haplotype with a nearby second indel (observed in the same sequencing reads) collectively led to an in-frame indel allele. In addition, six essential splice site-disrupting variants caused by indels resulted in no actual modification of the splice receptor or acceptor site motif. Overall,

annotation issues occurred in about 13% of the initial set, indicating that there is a need for annotation methods to take the underlying haplotype inferred from sequencing reads into account. We also analyzed the common putative *HLA-A* LoF frameshift variant rs1136691 in more detail, as *HLA-A* null alleles are known to occur very rarely in large transplantation databases (http://hla.alleles.org/alleles/nulls.html). An analysis of the sequencing reads corresponding to this variant revealed an alternative haplotype of several variants, and a Blast analysis of this sequence yielded a perfect match with an alternative unmapped contig of chromosome 6 (chr6_GL000256v2_alt in GRCh38). This observation suggests that the alternative haplotype may have been wrongly mapped to the closest sequence in the reference genome, resulting in an artefactual frameshift in the *HLA-A* gene. This hypothesis is consistent with the known genomic complexity and high level of polymorphism of the *HLA* region, which can lead to incorrect mapping (28). Finally, we noted that one variant, rs36046723 in the *ZNF852* gene, had an allele frequency >0.9999, suggesting that the reference genome carries the derived, low-frequency variant at this position. After curation, we retained 181 predicted LoF variants from 168 genes (**Table 1**).

### Characteristics of the LoF variants

The set of 181 predicted LoF variants defining the set of dispensable genes included premature stop-gains (40%), frameshifts (47%), and splice-site variants (13%; **Figure 1**). The variants could be classified further into those with a high or low predicted probability of being LoF (**Methods**): 27% presented features consistent with a low probability of LoF, whereas the remaining **73**% were predicted to have more severely damaging consequences and were therefore considered to have a high probability of LoF (**Figure 1** and **Supplementary Figure 3**). Despite possible differences in their impact, low- and high-probability LoF variants had similarly distributed global allele frequencies (**Supplementary Figure 4**; two-tailed Wilcoxon test *p*-value = 0.3115, based on ExAC allele frequencies; and 0.3341 based on GnomAD allele frequencies). Only 30 of the LoF variants finally retained were reported in at least one PubMed publication in the dbSNP database (29) (**Supplementary Table 1**). Manual inspection of these studies revealed that only 11 LoF variants had been experimentally demonstrated to abolish gene function (**Supplementary Table 1**). Focusing on the overlap between the 181 predicted LoF variants and the GWAS hits, we found that only two were reported in the GWAS catalog (30) as being significantly associated with a phenotypic trait: rs41272114

(*LPA*), associated with plasma plasminogen levels, lipoprotein A levels and coronary artery disease, and rs601338 (*FUT2*), associated with the levels of certain blood proteins, such as fibroblast growth factor 19. Finally, only 27 out of the 181 predicted LoF presented annotations in ClinVar, including 25 of them labelled as benign / likely benign, one as protective/risk factor (rs5744168 in *TLR5*, discussed in later sections of this manuscript), and another as pathogenic (rs17147990 in *HTN3*) (**Supplementary Table S1**). Pathogenicity for rs17147990 is only based on a publication reporting this mutation in the Histatin 3 gene (p.Tyr47Ter) (31), without any evidence of causality for a clinical phenotype. Overall, this analysis shows that most of the common LoF considered here present features consistent with severely damaging variants for gene function, although experimental characterization is largely lacking.

**Overlap of dispensable genes with disease-causing and essential genes**

We then explored the features characterizing the list of 168 putatively dispensable genes, and searched for those previously shown to be (i) associated with Mendelian diseases (*n*=3,622, OMIM (32)), or (ii) essential in human cell lines (*n*=1,920) or knockout mice (*n*=3,246) (33) (**Methods**). We focused on LoF variants predicted to be severely damaging, as described above (**Figure 1**). We found that three LoF variants from our list affected OMIM genes (*CLDN16, TMEM216, GUF1;* **Table 2**), whereas none affected essential genes (Fisher's exact *p*-value = 7.727e-06 and 6.249e-10 for human cell lines and knockout mice essential genes, respectively). It has been suggested that the common LoF variant rs760754693 of *CLDN16* — a gene associated with a renal disorder known as familial hypomagnesemia with hypercalciuria and nephrocalcinosis — affects the 5'-untranslated region of the gene rather than its coding sequence (34). A second methionine residue downstream from the affected position in *CLDN16* could potentially serve as the actual translation start site. The *TMEM216* variant rs10897158 is annotated as benign in ClinVar (35). *TMEM216* is required for the assembly and function of cilia, and pathogenic mutations of this gene cause Joubert, Meckel and related syndromes (36). The canonical transcript of *TMEM216* encodes a 145-amino acid protein. The rs10897158 splice variant (global frequency of homozygous individuals >70%) results in the synthesis of a longer protein (148 amino acids) corresponding to the most prevalent isoform (36), which could probably be considered to be the reference protein in humans. There is currently little evidence to support the third variant, rs141526764 (*GUF1*), having any pathogenic consequences. The association of *GUF1* with Mendelian

disease (early infantile epileptic encephalopathy) is reported in OMIM as provisional and based solely on the finding of a homozygous missense variant (A609S) in three siblings born to consanguineous parents (37). For all subsequent analyses, we filtered out the two variants considered highly likely not to be LoF (rs760754693, rs10897158), resulting in a final list of 179 predicted LoF variants corresponding to 166 genes (**Table 1**). The absence of common LoF variants predicted with a high degree of confidence in disease genes or essential genes is consistent with these genes being dispensable.

### Features of the set of putative dispensable protein-coding genes

We characterized the set of 166 putatively dispensable protein-coding genes further by performing a Gene Ontology (GO) enrichment analysis (**Methods**). We found a significant overrepresentation of genes involved in G-protein coupled receptor activity and related GO terms (Fisher's exact *p*-value= 5.50e-25; **Supplementary Table 2**). Such enrichment was driven by the presence of 41 olfactory receptor (OR) genes, accounting for 25% of the total set of dispensable genes, consistent with previous analyses (9). We then stratified dispensable genes according to their mutational damage, assessed by calculating the GDI score (38). The GO enrichment analyses conducted separately for the two sets gave very similar results, driven by the OR genes in both cases (**Supplementary Table 2**). After the removal of ORs, no additional functional categories were identified as displaying significant enrichment (**Supplementary Table 2**). Thus, a large number of protein-coding genes may be dispensable, but this has no apparent impact at the level of pathways or functional categories. Based on previous findings and the known specific features of OR genes (39, 40), we opted to consider the 41 OR and 125 non-OR dispensable genes separately in subsequent analyses (**Table 1**). In comparisons with a reference set of 382 OR, and 19,850 non-OR protein-coding genes (Methods), the coding lengths of the 41 OR (median = 945 nt) and the 125 non-OR (median = 1153.5 nt) dispensable genes were not significantly different from those of the corresponding non-dispensable OR (median = 945 nt; Wilcoxon test *p*-value= 0.6848) and non-OR (median = 1275 nt; Wilcoxon test *p*-value=0.1263) genes. The genomic distribution of dispensable OR genes displayed some clustering on some chromosomes, but did not differ significantly from that of the reference OR genes (**Supplementary Figure 5A**). The distribution of dispensable non-OR genes across autosomal chromosomes was also similar to that for the reference set, except that no dispensable genes were

present on the X and Y chromosomes (**Supplementary Figure 5B**). This finding suggests that common LoF variants on sexual chromosomes have been more efficiently purged from the population than autosomal variants, presumably because they have pathogenic effects in hemizygous males.

### Organ expression patterns of the dispensable protein-coding genes

We then investigated the expression patterns of the 166 putatively dispensable genes across organs and leukocyte types. For organs, we used RNA-seq expression data from the Illumina Body Map project (IBM) and the GTEX project, and for leukocytes, we used data from the BluePrint project (**Methods; Supplementary Table 3**). Most of the dispensable OR (30 of 34, 88%) were not found to be expressed in any of the datasets considered, consistent with the general expression pattern for all OR genes (328 of 355 OR genes not expressed, 92%; **Figure 2**). We found that 996 of a reference set of 17,948 non-OR protein-coding genes were not expressed in any of the databases considered (referred to hereafter as non-detected genes). Interestingly, the non-detected genes displayed a significant enrichment in dispensable genes relative to the reference set: odds ratio (OR): 3.34, 95% confidence interval (CI) 1.92-5.53, Fisher's exact test *p*-value 2.29e-05 **(Figure 2**). A similar pattern was observed for organ-specific genes, which were defined as genes expressed in <20% of the organs evaluated in the corresponding dataset: OR of 2.09 (CI 1.38-3.11) for IBM, *p*-value=3.67e-04, and OR of 3.56 (CI 2.41-5.22) for GTEX, *p*-value=1.53e-10. We further characterized the distribution of non-OR dispensable genes among the organ-specific genes for the various organs evaluated. Consistent with previous observations (11), the brain appeared to be the only organ in which the proportion of dispensable genes was significantly lower than that observed for organ-specific genes in both the IBM and the GTEX datasets: OR of 0.23 (CI 0.04-0.72), *p*-value=5.47e-03, and OR of 0.08 (CI 0.002-0.45), *p*-value=2.69e-04, respectively (**Supplementary Table 4**). Organ-pervasive genes were also found to be significantly depleted of dispensable genes, reflecting a lower degree of redundancy: OR of 0.09 (CI 0.04-0.18), *p*-value=6.47e-20 for IBM, and OR of 0.16 (CI 0.09-0.26), *p*-value=8.04e-19 for GTEX. The number of organs in which a gene is expressed was consistently, and negatively associated with the proportion of dispensable genes (linear regression *p*-values after adjustment for coding sequence

length < 2e-16 for both IBM and GTEX). Overall, genes widely expressed or specifically expressed in the brain are less dispensable than those with a more restricted pattern of expression.

### Expression patterns for dispensable genes in leukocytes

We then investigated the expression patterns of the 166 putatively dispensable genes in leukocytes. Human immune genes were classified on the basis of the RNA-seq data generated by the BluePrint project (41). We identified 7,323 adaptive leukocyte-expressed genes on the basis of their expression in B or T cells in at least 20% of the samples considered, and 9,039 innate leukocyte-expressed genes defined on the basis of their expression in macrophages, monocytes, neutrophils, or dendritic cells (DC) in at least 20% of the samples considered (**Supplementary Table 3**). We are aware that the main function of DCs is to present antigens to T cells, making their classification as "innate" both arbitrary and questionable. These leukocyte-expressed genes included no OR genes, and all the results therefore correspond to non-OR genes. A significant depletion of dispensable non-OR genes relative to the reference set was observed among the genes expressed in adaptive and innate leukocytes: OR of 0.20 (95% CI 0.10-0.34), *p*-value=8.80e-12, and OR of 0.20 (CI 0.12-0.33), *p*-value=9.98e-14, respectively; **Figure 2**). In total, 24 dispensable genes were identified as expressed in adaptive leukocytes (*n*=4 genes), innate leukocytes (*n*=10) or both (*n*=10). Detailed information about the common homozygous LoF variants of these genes is presented in **Supplementary Table 5.** Sixteen of these 24 genes had variants predicted to have highly damaging consequences, including a well-known stop-gain variant of *TLR5* (42). This truncating *TLR5* variant, which abolishes cellular responses to flagellin, appears to have evolved under neutrality (43). These genes also included *APOL3*, which is known to be involved in the response to infection with African trypanosomes (44). Thus, genes widely expressed in leukocytes are generally less dispensable than the reference set. However, specific immune-related genes may become LoF-tolerant due to functional redundancy (e.g. TLR5) or positive selection, by increasing protective immunity, for example. This aspect is considered in a subsequent section.

### Population distribution of the dispensable genes

We analyzed the distribution of the dispensable genes across the five specific populations considered: Africans (including African-Americans), East Asians, South Asians, Europeans (including Finnish), and Americans of Latino descent (**Figure 3).** Of the 125 non-OR genes, we found that 33 were dispensable in all populations, 16 in four populations, and 13 in three populations (homozygous LoF frequency >1% in each population). Conversely, 48 genes were population-specific, with Africans providing the largest fraction in both absolute (26 genes) and relative terms, as a reflection of their greater genetic diversity (45). The remaining 15 genes were found in two populations. Almost half the 41 dispensable OR were common to all five populations (*n*=19), 10 were present in four or three populations, two were present in two populations, and 10 were population-specific (including nine in Africans). As expected, the number of populations in which a gene was found to be dispensable correlated with the maximum frequency of homozygous individuals in the populations concerned (**Supplementary Figure 6**). Overall, 49% of the non-OR and 71% of the OR dispensable genes (>90% of the dispensable OR genes in non-African populations) were common to at least three human populations, suggesting a general process of tolerance to gene loss independent of the genomic background of the population.

**Negative selection of dispensable genes**

We then investigated the behavior of dispensable genes in terms of the functional scores associated with gene essentiality and selective constraints (**Figure 4**). We first evaluated a metric that was designed to assess the mutational damage amassed by a gene in the general population (GDI (38)) and three gene-level scores that measure the extent of recent and ongoing negative selection in humans (RVIS (46), pLI and pRec (12); **Methods**). Consistent with expectations, dispensable non-OR genes had much higher GDI (two-tailed Wilcoxon test *p*-value=6.52e-35), higher RVIS (*p*-value=1.30e-37), and lower pLI (*p*-value=1.548597e-22) values than non-dispensable non-OR genes **(Figure 4),** whereas no significant differences were found for pRec values (*p*-value=0.29). In analyses focusing on OR genes, GDI and RVIS values were also significantly higher for dispensable genes than for non-dispensable genes (*p*-values= 3.08e-06 and 1.96e-02, respectively), whereas no differences were found for the pLI and pRec distributions (*p*-value>0.05). When restricting the assessment of the GDI score to missense variants, the same trends where observed: p-value=1.24e-15 and 1.44e-02, for non-OR and OR comparisons, respectively. We then evaluated the strength of negative selection at a

deeper evolutionary level, using two inter-species conservation scores: the estimated proportion *f* of non-lethal non-synonymous mutations (47), which was obtained with SnIPRE, by comparing polymorphism within humans and divergence between humans and chimpanzee at synonymous and non-synonymous sites (48), and the GerpRS conservation score, obtained from alignments of sequences from multiple mammalian species (excluding humans) (49). Neither the *f* nor the GerpRS values obtained differed significantly (*p*-value >0.01) between dispensable and non-dispensable OR genes (**Figure 4**). However, dispensable non-OR genes had higher *f* and GerpRS values than non-dispensable non-OR genes (*p*-value= 1.51e-29 and = 9.70e-28, respectively), indicating that dispensable non-OR genes were more tolerant to non-synonymous variants than non-dispensable non-OR genes. Dispensable genes also had more human paralogs than other protein-coding genes (*p*-value= 8.76e-06), suggesting a higher degree of redundancy. For OR genes, the number of paralogs did not differ between dispensable and non-dispensable genes, further confirming that the negative selection parameters of dispensable and non-dispensable genes OR genes were similar. Overall, these results reveal a relaxation of the selective constraints acting at dispensable non-OR gene loci relative to non-dispensable genes, providing further evidence for evolutionary dispensability.

**Positive selection of common LoF variants**

We investigated the possibility that the higher frequency of some LoF mutations was due to a selective advantage conferred by gene loss (i.e., the “less-is-more” hypothesis), by searching for population-specific signatures of positive selection acting on these variants (17, 18). We considered two neutrality statistics: $F_{ST}$, which measures between-population differences in allele frequencies at a given locus (50), and integrated haplotype score (iHS) (51), which compares the extent of haplotype homozygosity around the ancestral and derived alleles in a given population. Both statistics could be computed for 68 variants fulfilling the quality control criteria (**Methods**). Considering a cutoff point of the 95$^{th}$ genome-wide percentile for each statistic (**Figure 5**), we detected 39 common LoF alleles in putatively dispensable genes displaying signals suggestive of positive selection, as attested by their low iHS (*n*=7), high $F_{ST}$ values (*n*=24), or both (*n*=8; **Supplementary Table 6**). Seven of these variants in OR genes had only high $F_{ST}$ values, suggestive of genetic drift related to the ongoing pseudogenization of ORs (40). We also noted that the LoF mutation (rs2039381) of *IFNE* displayed significant levels of population differentiation (e.g. $F_{ST}$=0.25 for GIH vs. CEU, $P_{emp}$ = 0.002). This result

is intriguing as *IFNE* encodes IFNε, which plays an important role in protective immunity against microbes in the female reproductive tract in mice (52, 53). This nonsense variant (Q71X) is predicted to decrease the length of the encoded protein by two thirds, but this has not been validated experimentally. The proportion of homozygotes is highest in East Asia (3.5%) and South Asia (2%), is much lower in Europe (0.02%), and does not differ between males and females. The iHS scores were not significant for this variant (**Supplementary Figure 7**), and neither were other selection scores, such as Tajima's *D*, Fu and Li's *D** and *F* *, and Fay and Wu's *H*, previously obtained in a large evolutionary genetic study of human interferons (54). In light of these observations, the most likely explanation for the high $F_{ST}$ observed at this locus is genetic drift.

**LoF mutations of *FUT2* and *APOL3* are under positive selection**

Eight common LoF variants provided more robust evidence for positive selection, as they had both a high $F_{ST}$ and a low iHS (**Supplementary Table 6**, **Figure 5**). For five of these variants, there was no obvious relationship between the gene concerned and a possible selective advantage. However, one variant with a high $F_{ST}$ ($F_{ST}$ = 0.25 for ITU vs. CHB, $P_{emp}$= 0.012) and a low iHS ( iHS -2.33, $P_{emp}$ = 0.004 in CHS) was located in *SLC22A14,* which has been shown to be involved in sperm motility and male infertility in mice (55). Finally, the two remaining variants were in the *FUT2* and *APOL3* genes, which are known to be involved in defense against infections. Consistent with previous observations, we observed high $F_{ST}$ values ($F_{ST}$ = 0.54 for CEU vs. CHS, $P_{emp}$= 0.002) and extended haplotype homozygosity iHS =-1.7, $P_{emp}$ = 0.03 in CEU) around the LoF mutation (rs601338) in *FUT2* (**Supplementary Figure 7**). This gene is involved in antigen production in the intestinal mucosa, and null variants are known to lead to the non-secretor phenotype conferring protection against common enteric viruses, such as norovirus (25, 56), and rotavirus (57, 58). We also identified a novel hit at the *APOL3* LoF variant rs11089781. This nonsense variant (Q58X) was detected only in African populations (15-33% frequency), in which it had a low iHS (iHS=-2.75 in MSL, $P_{emp}$ = 0.001), indicating extended haplotype homozygosity around the LoF mutation in these populations (**Supplementary Figure 7**). *APOL3* is located within a cluster of *APOL* genes including *APOL1*. These two members of the six-member APOL cluster, APOL1 and APOL3, are known to be involved in defense against African trypanosomes (59, 44). These analyses indicate that, although positive selection remains rare in

humans, it may have increased the frequency of LoF variants when gene loss represents a selective advantage.

## Discussion

We identified 166 putatively dispensable human protein-coding genes. Even after manual curation, it is likely that a certain proportion of the variants retained for these genes are not actually LoF, and the availability of experimental validation for only a small fraction of the predicted LoF variants is one of the limitations of this study. It is also likely that additional dispensable genes could be detected in the general population on the basis of common homozygosity for other types of genetic variants abrogating protein function, notably those for which whole-exome sequencing is not particularly suitable, such as deletions of more than 50 nucleotides (60). The putatively dispensable human protein-coding genes identified included 120 that overlapped either with the total list of 2,641 genes apparently tolerant to homozygous rare LoF reported from bottlenecked or consanguineous populations (10, 11, 13, 14) or with the list of 253 genes initially identified from individuals of the 1000 Genomes Project (9) (**Supplementary Figure 8**), or both. These partly discordant results are probably due to differences in study designs, in particular in terms of sample sizes, population structures, and frequencies of the LoF variants studied. Most of the 46 genes specific to our study had a maximum homozygote frequency below 0.05 or had a higher frequency in only one or two of the five studied populations, as for *FUT2* and *IFNE*. Our set of dispensable genes was strongly enriched in OR genes, as previously reported (9, 61). However, dispensable OR genes had no particular features distinguishing them from non-dispensable OR genes other than slightly higher GDI and RVIS values. This finding is consistent with the notion that the number of functional ORs has decreased during evolution in humans (62), and provides evidence for ongoing pseudogenization (63, 40). Conversely, dispensable non-OR genes displayed a strong relaxation of selective constraints relative to non-OR non-dispensable genes at both the inter-species and intra-species levels. In addition, the set of dispensable non-OR genes was depleted of genes widely expressed in the panel of organs evaluated, brain-specific genes and genes expressed in leukocytes. This suggests that the redundancy observed for some microbial sensors and effectors (64, 18) does not necessarily translate into higher rates of gene dispensability.

The set of dispensable genes identified here probably largely corresponds to genes undergoing pseudogenization (65, 66) due to present-day superfluous molecular function at the cell,

organ or organism levels, or a redundant function in the genome that may be recovered (e.g. by paralogous genes or alternative pathways). This is consistent with the observation that most of the dispensable genes were common to at least three human populations. The overlap was particularly large for OR genes, which are strongly enriched in dispensable genes, and generally present a strong relaxation of selective constraints and signs of ongoing pseudogenization (40). Another example is provided by *TLR5,* encoding a cell-surface receptor for bacterial flagellin, which harbors a dominant negative stop mutation at high population frequencies (43, 42, 67). This finding is consistent with the notion that a substantial proportion of modern-day humans can survive with complete TLR5 deficiency (42). These observations also suggest that additional mechanisms of flagellin recognition, such as those involving the NAIP-NLRC4 inflammasome (68, 69), may provide sufficient protection in the absence of TLR5. Finally, 45 of our dispensable genes belong to the set of 2,278 Ensembl/GENCODE coding genes recently reported to display features atypical of protein-coding genes (70). Our study may therefore provide additional candidates for inclusion in the list of potential non protein-coding genes.

High population frequencies of LoF variants may also reflect recent and ongoing processes of positive selection favoring gene loss (i.e. the “less-is-more” paradigm) (17). Two of the eight variants with the most robust signals of positive selection were located in genes involved in resistance to infectious diseases. The *FUT2* gene is a well-known example of a gene for which loss is an advantage, as it confers Mendelian resistance to common enteric viruses and has a profile consistent with positive selection. However, non-secretor status has also been associated with predisposition to Crohn’s disease (71), Behçet’s disease (72), and various bacterial infections (73), including otitis media (74), suggesting an advantage for secretor status. Accordingly, an evolutionary genetics study concluded that *FUT2* genetic diversity was compatible with the action of both positive and balancing selection on secretor status (75). An interesting new finding from this study is provided by the *APOL3* LoF variant (rs11089781), which we found to display signals of recent, positive selection in Africans. APOL3 and APOL1 are known to be involved in the response to African trypanosomes (44). Two variants encoding APOL1 proteins with enhanced trypanolytic activity are present only in African populations, in which they harbor signatures of positive selection despite increasing the risk of kidney disease (76). Interestingly, the *APOL3* LoF variant was also recently associated with nephropathy independently of the effect of the two late-onset kidney disease-risk *APOL1* variants, which are not in strong linkage

disequilibrium with rs11089781 (77). A physical interaction occurs between APOL1 and APOL3 (77), and APOL1 may protect against pathogens more effectively when not bound to APOL3. Similar mechanisms may, therefore, be involved in the positive selection of the *APOL1* kidney disease-risk alleles and the *APOL3* LoF variant in African populations. Additional common LoF homozygotes could probably be further identified in other unstudied populations, or involving variants not currently predicted to be LoF *in silico*. Improvements in the high-confidence identification of dispensable genes will make it possible to identify biological functions and mechanisms that are, at least nowadays, redundant, or possibly even advantageous, for human survival.

**Acknowledgments**

We thank the Laboratory of Clinical Bioinformatics and both branches of the Laboratory of Human Genetics of Infectious Diseases, Yuval Itan, Sophie Saunier and Corinne Antignac for helpful discussions and support. The Laboratory of Clinical Bioinformatics was supported by the French National Research Agency (Agence Nationale de la Recherche, ANR) “Investissements d’Avenir” program, ANR-10-IAHU-01 and the Christian Dior Couture, Dior. The Laboratory of Human Genetics of Infectious Diseases was supported in part by grants from the French National Agency for Research (ANR) under the “Investissement d’avenir” program (grant number ANR-10-IAHU-01), the Fondation pour la Recherche Médicale (Equipe FRM EQU201903007798), the St. Giles Foundation, and the Rockefeller University. The laboratory of L.Q.-M. is supported by the Institut Pasteur, the Collège de France, the French Government’s Investissement d’Avenir program, Laboratoires d’Excellence “Integrative Biology of Emerging Infectious Diseases” (ANR-10- LABX-62-IBEID) and “Milieu Intérieur” (ANR-10-LABX-69-22701), and the Fondation pour la Recherche Médicale (Equipe FRM DEQ20180339214). David Cooper and Peter Stenson acknowledge the financial support of Qiagen Inc through a License Agreement with Cardiff University.

## Methods

### Exome sequencing data

Human genetic variants from the Exome Aggregation Consortium (ExAC) database (12) (http://exac.broadinstitute.org/) were downloaded on September 9th 2016, release 0.3.1, non-TCGA subset. Variants from Exome data of the Genome Aggregation Database (gnomAD, http://gnomad.broadinstitute.org/) were obtained on February 28th 2017, release 2.0.1. For consistency with ExAC and GnomAD pipelines, variants were annotated with the Ensembl Variant Effect Predictor VEP (78) (v81 for ExAC, v85 for gnomAD), with loss-of-function (LoF) annotations from LOFTEE ((12) https://github.com/konradjk/loftee). *Homo sapiens* genome build GRCh37/hg19 was used with both databases. Analyses were restricted throughout this study to a background set of 20232 human protein-coding genes obtained from BioMart Ensembl 75, version Feb 2014 (GRCh37.p13; (79)). Protein-coding genes were defined as those containing an open reading frame (ORF). By contrast, pseudogenes were typically defined as gene losses resulting from fixations of null alleles that occurred in the human lineage after a speciation event, some of them may actually be human-specific, i.e. fixed after the human-chimpanzee divergence. However, the definition might be larger, including the so-called processed pseudogenes, corresponding to DNA sequences reverse-transcribed from RNA and randomly inserted into the genome (65, 66). A total of 20,232 protein-coding genes and 13,921 pseudogenes were reported by Ensembl following the Ensembl Genebuild workflow incorporating the HAVANA group manual annotations (26). Among the protein-coding genes, 382 genes were identified with a gene name starting with “Olfactory receptor” (**Supplementary Table 7**).

VEP annotations were done against the set of Ensembl Transcript IDs associated to Ensembl protein coding genes, focusing on the canonical transcript as described in the ExAC Flagship paper (Lek et al.). This was done through VEP option “--canonical”, which provides the variant consequence for what is considered to be the “canonical transcript” of a gene. As reported in (80) and further detailed in Ensembl documentation (http://www.ensembl.org/info/website/glossary.html) for human, the canonical transcript for a gene is set according to the following hierarchy: (1) Longest CCDS translation with no stop codons. (2) If no (1), choose the longest Ensembl/Havana merged translation with no stop codons. (3). If no (2), choose the longest translation with no stop codons. (4). If no translation, choose the longest non-protein-coding transcript. However, as acknowledged in the Ensembl documentation, the canonical transcript does not necessarily reflect the most biologically relevant transcript of a gene. For the purpose of this study, we thus further focused on variants affecting canonical transcripts when these are considered in turn the “principal isoform” of the associated gene, i.e. its most functionally important transcript. To that aim, we used transcript annotations from the APPRIS database (28) (downloaded on downloaded March 2nd 2017, using Gencode19/Ensembl74). APPRIS is a system to annotate principal isoforms based on a range of computational methods evaluating structural information, presence of functionally important residues and conservation evidence from cross-species alignments.

**Loss-of-function variants and definition of the set of dispensable genes**

Loss-of-function (LoF) variants were considered here as those predicted to lead to an early stop-gain, indel frameshift or essential splice-site disruption (i.e, splice-site donor and splice-site acceptor variants). Following the criteria used in the ExAC flagship paper (12), only variants with a genotype quality ≥ 20, depth ≥ 10, a call rate > 80%, mapped to canonical isoforms and labeled as "high-confidence" LoF variants by LOFTEE, were retained. LOFTEE ((12), https://github.com/konradjk/loftee) is a VEP plugin allowing to flag and filter variants according to quality control criteria characteristic of falsely considered LoF variants. Thus, the following low-confidence LoF variants flagged by LOFTEE were filtered out: variants for which the purported LoF allele is the ancestral state (across primates), stop-gain and frameshift variants in the last 5% of the transcript, or in an exon with non-canonical splice sites around it (i.e. intron does not start with GT and end with AG), and splice-site variants in small introns (<15 bp), in an intron with a non-canonical splice site or rescued by nearby in-frame splice sites. In a previous work we showed that features associated to low-confidence loss-of-function variants are enriched in common (AF>5%) and in homozygous LoF in the general population (5). Consistent with such observation, putatively false LoF variants filtered by LOFTEE are enriched in common variants (81), and would confound the detection of dispensable genes as defined above unless they are filtered out. Following https://macarthurlab.org/2016/03/17/reproduce-all-the-figures-a-users-guide-to-exac-part-2/, variants in the top 10 most-multiallelic kilobases of the human genome were filtered out, i.e: Chr14:106330000-106331000; Chr2:89160000-89161000; Chr14:106329000-106330000; Chr14:107178000-107179000; Chr17:18967000-18968000; Chr22:23223000-23224000; Chr1:152975000-152976000; Chr2:89161000-89162000; Chr14:107179000-107180000; Chr17:19091000-19092000. For GnomAD variants, a heterozygote genotype allele balance > 0.2 was further required. In addition to the previous criteria, a set of filters for LoF variants was adopted in this work, retaining variants with a variant quality score recalibration (VQSR) equal to 'PASS' both for single nucleotide polymorphisms (SNPs) and frameshifts and affecting the APPRIS principal isoform (27) as described above. Dispensable genes were defined as those presenting a LoF variant with a frequency of homozygous individuals higher than 1% in at least one of the 5 main populations considered, i.e.: Africans (including African Americans), East Asians, South Asians, Europeans (Finnish and Non-Finnish), and Americans.

**Impact prediction of LoF variants**

The mRNA region capable of triggering transcript degradation by NMD upon an early stop-gain was defined as in a previous study (5)**,** according to HAVANA annotation guidelines (v.20). Specifically, the NMD-target region of a transcript was defined as those positions more than 50 nucleotides upstream from the 3'-most exon-exon junction. Transcripts bearing stop-gain variants in these regions are predicted to be degraded by NMD (82). Molecular impact prediction of splice-disrupting variants was performed with Human Splicing Finder (HSF) software ((83); online version 3.1 available at http://www.umd.be/HSF/ with default parameters). HSF classifies splicing variants into five categories: unknown impact, no impact, probably no impact, potential alternation, most probably affecting variant. LoF variants were classified into those predicted to be LoF with low and high probability. Low-

probability LoF arbitrarily include (i) stop-gains and frameshift variants truncating the last 15% of the protein sequence, which might translate into small truncations in the final protein, or mapping into the first 100 nucleotides of the transcript, which has been reported as a window length in which LoF variants are often recovered by means of alternative start site usage (84); and (ii) essential splice site variants with an unknown or low computationally predicted impact on splicing motifs based on position weight matrices, maximum entropy and motif comparison methods (Methods). High-probability LoF variants include (i) stop-gains and frameshifts truncating more than 15% of the protein sequence or occurring in a region prone to transcript degradation by NMD, which probably result in the complete abolition of protein production (Methods); and (ii) putative splice-site variants with an intermediate or high computationally predicted impact (Methods). All associations of LoF variants reported in the GWAS catalog (version v1.0, date 2019-01-11) were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/gwas/. As for Clinvar annotations, the 2020/03/02 release was used (35).

**Gene Ontology enrichment analysis**

Gene Ontology enrichment analysis was performed with DAVID functional annotation tool (85). Only terms with a Benjamini-corrected Fisher's exact test *p*-value < 0.05 were retained.

**Gene expression patterns across human organs and immune cell types**

Two lists of organ and tissue-expressed genes were defined on the basis of the RNA-seq expression data from the Illumina Body Map project (IBM, 15,688 expressed genes), and the GTEX project (16,762 expressed genes). First, RNA-seq expression data from a panel of 11 human organs and tissues (one sample each) from the Illumina Body Map project (IBM) were extracted from the Expression Atlas database (EBI accession E-MTAB-513; 1 February 2017 release; https://www.ebi.ac.uk/gxa/experiments/E-MTAB-513/Results). The list of organs and tissues included: adipose tissue, brain, breast, colon, heart, kidney, liver, lung, ovary, skeletal muscle tissue, and testis. It should be noted that leukocyte, lymph node, adrenal, prostate and thyroid gland data were removed from these datasets. A total of 15,688 organ-expressed genes were defined as being expressed in more than 3 transcripts per million (TPM; according to (86)) in at least one of the IBM samples considered. Second, RNA-seq expression data from a panel of 24 human organs (multiple samples per organ) from the GTEX project (87) were extracted from https://gtexportal.org/home/tissueSummaryPage (version V6p). Blood, blood vessel, salivary gland, adrenal gland, thyroid, pituitary and bone marrow data were removed from these datasets. A total of 16,762 organ-expressed genes were defined as being expressed in more than 3 TPM in at least 20% of the samples from at least one GTEX organ. The organ-expressed set of genes previously defined was further classified into a set of "organ-specific genes" and "organ-pervasive genes", depending on whether the gene was defined as expressed in <20% (organ-specific) or >80% of the organs and tissue types evaluated in the corresponding dataset (i.e. IBM or GTEX).

In addition, the RNA-seq dataset generated by the BluePrint project (41) was used as a reference set for gene expression for the different immune cell types from human venous blood (August 19, 2017 release, data available at http://dcc.blueprint-epigenome.eu/#/files). Only cell types with more than two samples were considered. In total, 85 libraries were retained, including: 9 B-cell and 17 T-cell samples

(collectively considered as adaptive immune cell types), and 15 monocyte, 25 macrophage, 6 dendritic cell and 13 neutrophil samples (collectively considered as innate immune cell types). A total of 7,346 adaptive immune cell-expressed genes were detected as those expressed in more than 3 TPM in B cells or T cells in at least in 20% of the corresponding samples collectively considered. Similarly, 9,069 innate immune cell-expressed genes were defined here as expressed in more than 3 TPM in macrophages, monocytes, neutrophils, or dendritic cells in at least 20% of the corresponding samples collectively considered. Full details about the libraries used, as provided by the BluePrint project, are reported in **Supplementary Table 8**. The set of genes not found to be expressed in any of the previous lists was determined from the complement of the reference list of 20120 protein-coding genes defined as Ensembl Biomart, release 75 (79).

**Gene-level annotations**

The following gene-level features associated with natural selection were obtained: gene damage index (GDI) scores, a gene-level metric of the mutational damage that has accumulated in the general population, based on CADD scores, were taken from (38). High GDI values reflect highly damaged genes. The residual variation intolerance score (RVIS percentile, provided in (46)), assesses the gene's departure from the mean number of common functional mutations in genes with a similar mutational burden in humans. High RVIS percentiles reflect genes that are highly tolerant to variation. The gene probability of loss-of-function intolerance (pLI, (12)), estimating the depletion of rare and *de novo* protein-truncating variants relative to expectations derived from a neutral model of *de novo* variation on ExAC exomes data, and pRec, estimating gene intolerance to two rare and *de novo* protein-truncating variants (analogous to recessive genes) were obtained from the ExAC Browser (release 0.3.1, (12)). pLI and pRec values close to 1 indicate gene intolerance to heterozygous and homozygous loss-of-function and to homozygous mutations, respectively. "*f*" values were obtained through SnIPRE (48) as described in (47), based on a comparison of polymorphism and divergence at synonymous and non-synonymous sites. GerpRS scores (49) were downloaded from http://mendel.stanford.edu/SidowLab/downloads/gerp/. Data on the number of human paralogs for each gene were collected from the OGEE database (88). Monogenic Mendelian disease genes were obtained as described in Chong et al. (89): OMIM raw data files were downloaded from (90). Phenotype descriptions containing the word 'somatic' were flagged as 'somatic', and those containing 'risk', 'quantitative trait locus', 'QTL', '{', '[' or 'susceptibility to' were flagged as 'complex'. Monogenic Mendelian genes were defined as those having a supporting evidence level of 3 (i.e. the molecular basis of the disease is known) and not having a 'somatic' or 'complex' flag. (**Supplementary Table 7**). Genes essential in human cell lines and in knockout mice were obtained from (33) (**Supplementary Table 7**).

**Genome-wide scans for recent positive selection at loss-of-function mutations**

We tested for the occurrence of positive selection of loss-of-function mutations, by calculating two neutrality statistics: the interpopulation $F_{ST}$, which identifies loci displaying high levels of variation in

allele frequencies between groups of populations (50), and the intrapopulation integrated haplotype score (iHS) (51), which compares the extent of haplotype homozygosity at the ancestral and derived alleles. Positive selection analyses were confined to biallelic SNPs found in the 1000 Genomes Project phase 3 data (91), including 2,504 individuals from 26 populations, assigned to five meta-populations and predicted to have severely damaging consequences. (**Supplementary Table 9**). Multiallelic SNPs, SNPs not detected in the 1000 Genomes Project and indel frameshifts were discarded in the positive selection analysis. For $F_{ST}$ calculation, we investigated a total of 75 LoF mutations that passed quality filters, and compared the allele frequencies of these variants in 26 populations, to the allele frequencies of the same mutations in the European (CEU) and African (YRI) reference populations (**Supplementary Table 9**). More specifically, we compared allele frequencies in populations from the African (AFR), East Asian (EAS) and South Asian (SAS) meta-populations to allele frequencies in the CEU groups, and allele frequencies in populations from the American (AMR) and European (EUR) meta-populations to allele frequencies in the YRI group. For the detection of candidate variants for positive selection based on $F_{ST}$ values, we used an outlier approach and considered LoF mutations presenting $F_{ST}$ values located in the top 5% of the distribution of $F_{ST}$ genome-wide. We identified 32 LoF mutations presenting high $F_{ST}$ values in at least one population, in 32 genes (including 8 mutations located in OR genes). For haplotype-based iHS score calculations, we first defined the derived allele state of each SNP based on the 6-EPO alignment, and retained only SNPs with a derived allele frequency (DAF) between 10% and 90% to maximize the power of iHS to detect selective signals. These additional filters led to a total of 68 LoF mutations to be investigated. We calculated iHS scores in 100 kb windows with custom-generated scripts and normalized values. For the detection of selection events targeting derived alleles, we considered LoF mutations located in the top 5% most negative iHS values genome-wide and found a total of 15 LoF mutations with iHS scores in the lowest 5% of t iHS values genome-wide in at least one population (including 1 mutation located in an OR gene).

**Data and Materials Availability**

All data used in the paper are present in the main text and SI appendix.

## References


1. F. S. Alkuraya, Human knockout research: new horizons and opportunities. *Trends Genet.* **31**, 108–115 (2015).
2. V. M. Narasimhan, Y. Xue, C. Tyler-Smith, Human Knockout Carriers: Dead, Diseased, Healthy, or Improved? *Trends Mol. Med.* **22**, 341–351 (2016).
3. J.-L. Casanova, L. Abel, Human genetics of infectious diseases: Unique insights into immunological redundancy. *Semin. Immunol.* **36**, 1–12 (2018).
4. G. Courtois, *et al.*, A hypermorphic IkappaBalpha mutation is associated with autosomal dominant anhidrotic ectodermal dysplasia and T cell immunodeficiency. *J. Clin. Invest.* **112**, 1108–1115 (2003).
5. A. Rausell, *et al.*, Analysis of Stop-Gain and Frameshift Variants in Human Innate Immunity Genes. *PLoS Comput Biol* **10**, e1003757 (2014).
6. M. K. Vickaryous, B. K. Hall, Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev.* **81**, 425 (2006).
7. P. D. Stenson, *et al.*, The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **136**, 665–677 (2017).
8. C. Picard, *et al.*, International Union of Immunological Societies: 2017 Primary Immunodeficiency Diseases Committee Report on Inborn Errors of Immunity. *J. Clin. Immunol.* **38**, 96–128 (2018).
9. D. G. MacArthur, *et al.*, A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* **335**, 823–828 (2012).
10. E. T. Lim, *et al.*, Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. *PLoS Genet.* **10**, e1004494 (2014).
11. P. Sulem, *et al.*, Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47**, 448–452 (2015).
12. M. Lek, *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
13. V. M. Narasimhan, *et al.*, Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474–477 (2016).
14. D. Saleheen, *et al.*, Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–239 (2017).
15. M. A. Nowak, M. C. Boerlijst, J. Cooke, J. M. Smith, Evolution of genetic redundancy. *Nature* **388**, 167–171 (1997).
16. M. V. Olson, When Less Is More: Gene Loss as an Engine of Evolutionary Change. *Am. J. Hum. Genet.* **64**, 18–23 (1999).
17. L. B. Barreiro, L. Quintana-Murci, From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat. Rev. Genet.* **11**, 17–30 (2010).
18. L. Quintana-Murci, A. G. Clark, Population genetic tools for dissecting innate immunity in humans. *Nat. Rev. Immunol.* **13**, 280–293 (2013).
19. L. H. Miller, S. J. Mason, D. F. Clyde, M. H. McGinniss, The Resistance Factor to *Plasmodium vivax* in Blacks: The Duffy-Blood-Group Genotype, *FyFy*. *N. Engl. J. Med.* **295**, 302–304 (1976).
20. C. Tournamille, Y. Colin, J. P. Cartron, C. Le Van Kim, Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy–negative individuals. *Nat. Genet.* **10**, 224–228 (1995).
21. M. Samson, *et al.*, Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* **382**, 722–725 (1996).
22. R. Liu, *et al.*, Homozygous Defect in HIV-1 Coreceptor Accounts for Resistance of Some Multiply-Exposed Individuals to HIV-1 Infection. *Cell* **86**, 367–377 (1996).
23. M. Dean, *et al.*, Genetic Restriction of HIV-1 Infection and Progression to AIDS by a Deletion Allele of the CKR5 Structural Gene. *Science* **273**, 1856–1862 (1996).
24. L. Lindesmith, *et al.*, Human susceptibility and resistance to Norwalk virus infection. *Nat. Med.* **9**, 548–553 (2003).
25. M. Thorven, *et al.*, A Homozygous Nonsense Mutation (428G->A) in the Human Secretor (FUT2) Gene Provides Resistance to Symptomatic Norovirus (GGII) Infections. *J. Virol.* **79**, 15351–15355 (2005).
26. B. L. Aken, *et al.*, The Ensembl gene annotation system. *Database* **2016**, baw093 (2016).
27. J. M. Rodriguez, *et al.*, APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* **41**, D110-117 (2013).

28. D. Y. C. Brandt, *et al.*, Mapping Bias Overestimates Reference Allele Frequencies at the *HLA* Genes in the 1000 Genomes Project Phase I Data. *G3amp58 GenesGenomesGenetics* **5**, 931–941 (2015).
29. S. T. Sherry, *et al.*, dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
30. A. Buniello, *et al.*, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
31. L. M. Sabatini, E. A. Azen, Two coding change mutations in the HIS2(2) allele characterize the salivary histatin 3-2 protein variant. *Hum. Mutat.* **4**, 12–19 (1994).
32. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), Online Mendelian Inheritance in Man, OMIM®, https://omim.org/ (2018).
33. I. Bartha, J. di Iulio, J. C. Venter, A. Telenti, Human gene essentiality. *Nat. Rev. Genet.* **19**, 51–62 (2017).
34. S. Weber, *et al.*, Novel paracellin-1 mutations in 25 families with familial hypomagnesemia with hypercalciuria and nephrocalcinosis. *J. Am. Soc. Nephrol. JASN* **12**, 1872–1881 (2001).
35. M. J. Landrum, *et al.*, ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
36. E. M. Valente, *et al.*, Mutations in TMEM216 perturb ciliogenesis and cause Joubert, Meckel and related syndromes. *Nat. Genet.* **42**, 619–625 (2010).
37. A. A. Alfaiz, *et al.*, West syndrome caused by homozygous variant in the evolutionary conserved gene encoding the mitochondrial elongation factor GUF1. *Eur. J. Hum. Genet. EJHG* **24**, 1001–1008 (2016).
38. Y. Itan, *et al.*, The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc. Natl. Acad. Sci.* **112**, 13615–13620 (2015).
39. Y. Gilad, D. Lancet, Population differences in the human functional olfactory repertoire. *Mol. Biol. Evol.* **20**, 307–314 (2003).
40. D. Pierron, N. G. Cortés, T. Letellier, L. I. Grossman, Current relaxation of selection on the human genome: tolerance of deleterious mutations on olfactory receptors. *Mol. Phylogenet. Evol.* **66**, 558–564 (2013).
41. H. G. Stunnenberg, *et al.*, The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1145–1149 (2016).
42. J.-L. Casanova, L. Abel, L. Quintana-Murci, Human TLRs and IL-1Rs in Host Defense: Natural Insights from Evolutionary, Epidemiological, and Clinical Genetics. *Annu. Rev. Immunol.* **29**, 447–491 (2011).
43. L. B. Barreiro, *et al.*, Evolutionary Dynamics of Human Toll-Like Receptors and Their Different Contributions to Host Defense. *PLoS Genet.* **5**, e1000562 (2009).
44. F. Fontaine, *et al.*, APOLs with low pH dependence can kill all African trypanosomes. *Nat. Microbiol.* **2**, 1500–1506 (2017).
45. R. Nielsen, *et al.*, Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
46. S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, D. B. Goldstein, Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet* **9**, e1003709 (2013).
47. M. Deschamps, *et al.*, Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. *Am. J. Hum. Genet.* **98**, 5–21 (2016).
48. K. E. Eilertson, J. G. Booth, C. D. Bustamante, SnIPRE: selection inference using a Poisson random effects model. *PLoS Comput. Biol.* **8**, e1002806 (2012).
49. E. V. Davydov, *et al.*, Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
50. K. E. Holsinger, B. S. Weir, Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nat. Rev. Genet.* **10**, 639–650 (2009).
51. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
52. K. Y. Fung, *et al.*, Interferon- Protects the Female Reproductive Tract from Viral and Bacterial Infection. *Science* **339**, 1088–1092 (2013).
53. S. A. Stifter, *et al.*, Defining the distinct, intrinsic properties of the novel type I interferon, IFNϵ. *J. Biol. Chem.* **293**, 3168–3179 (2018).
54. J. Manry, *et al.*, Evolutionary Genetic Dissection of Human Interferons. *J. Exp. Med.* **208**, 2747–2759 (2011).
55. S. Maruyama, *et al.*, A critical role of solute carrier 22a14 in sperm motility and male fertility in mice. *Sci. Rep.* **6**, 36468 (2016).
56. L. Lindesmith, *et al.*, Human susceptibility and resistance to Norwalk virus infection. *Nat. Med.* **9**, 548–553 (2003).
57. J. Nordgren, *et al.*, Both Lewis and Secretor Status Mediate Susceptibility to Rotavirus Infections in a Rotavirus Genotype–Dependent Manner. *Clin. Infect. Dis.* **59**, 1567–1573 (2014).

58. D. C. Payne, *et al.*, Epidemiologic Association Between *FUT2* Secretor Status and Severe Rotavirus Gastroenteritis in Children in the United States. *JAMA Pediatr.* **169**, 1040 (2015).

59. L. Vanhamme, *et al.*, Apolipoprotein L-I is the trypanosome lytic factor of human serum. *Nature* **422**, 83–87 (2003).

60. D. Shigemizu, *et al.*, IMSindel: An accurate intermediate-size indel detection tool incorporating de novo assembly and gapped global-local alignment with split read analysis. *Sci. Rep.* **8**, 5608 (2018).

61. A. B. Alsalem, A. S. Halees, S. Anazi, S. Alshamekh, F. S. Alkuraya, Autozygome Sequencing Expands the Horizon of Human Knockout Research and Provides Novel Insights into Human Phenotypic Variation. *PLoS Genet.* **9**, e1004030 (2013).

62. M. Nei, Y. Niimura, M. Nozawa, The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Nat. Rev. Genet.* **9**, 951–963 (2008).

63. S. Alonso, S. Lopez, N. Izagirre, C. de la Rua, Overdominance in the Human Genome and Olfactory Receptor Activity. *Mol. Biol. Evol.* **25**, 997–1001 (2008).

64. S. Nish, R. Medzhitov, Host Defense Pathways: Role of Redundancy and Compensation in Infectious Disease Phenotypes. *Immunity* **34**, 629–636 (2011).

65. X. Wang, W. E. Grus, J. Zhang, Gene Losses during Human Origins. *PLoS Biol.* **4**, e52 (2006).

66. H. L. Kim, T. Igawa, A. Kawashima, Y. Satta, N. Takahata, Divergence, demography and gene loss along the human lineage. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 2451–2457 (2010).

67. H. Quach, *et al.*, Different selective pressures shape the evolution of Toll-like receptors in human and African great ape populations. *Hum. Mol. Genet.* **22**, 4829–4840 (2013).

68. B. Ferwerda, *et al.*, Human Dectin-1 Deficiency and Mucocutaneous Fungal Infections. *N. Engl. J. Med.* **361**, 1760–1767 (2009).

69. Y. Zhao, F. Shao, The NAIP-NLRC4 inflammasome in innate immune detection of bacterial flagellin and type III secretion apparatus. *Immunol. Rev.* **265**, 85–102 (2015).

70. F. Abascal, *et al.*, Loose ends: almost one in five human genes still have unresolved coding status. *Nucleic Acids Res.* **46**, 7070–7084 (2018).

71. D. P. B. McGovern, *et al.*, Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Hum. Mol. Genet.* **19**, 3468–3476 (2010).

72. M. Takeuchi, *et al.*, Dense genotyping of immune-related loci implicates host responses to microbial exposure in Behçet's disease susceptibility. *Nat. Genet.* **49**, 438–443 (2017).

73. J. Le Pendu, N. Ruvoën-Clouet, E. Kindberg, L. Svensson, Mendelian resistance to human norovirus infections. *Semin. Immunol.* **18**, 375–386 (2006).

74. R. L. P. Santos-Cortez, *et al.*, FUT2 Variants Confer Susceptibility to Familial Otitis Media. *Am. J. Hum. Genet.* **103**, 679–690 (2018).

75. A. Ferrer-Admetlla, *et al.*, A natural history of FUT2 polymorphism in humans. *Mol. Biol. Evol.* **26**, 1993–2003 (2009).

76. G. Genovese, *et al.*, Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**, 841–845 (2010).

77. K. L. Skorecki, *et al.*, A null variant in the apolipoprotein L3 gene is associated with non-diabetic nephropathy. *Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. - Eur. Ren. Assoc.* **33**, 323–330 (2018).

78. W. McLaren, *et al.*, The Ensembl Variant Effect Predictor. *Genome Biol.* **17** (2016).

79. D. R. Zerbino, *et al.*, Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).

80. T. J. P. Hubbard, *et al.*, Ensembl 2009. *Nucleic Acids Res.* **37**, D690-697 (2009).

81. K. J. Karczewski, *et al.*, Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* (2019) https:/doi.org/10.1101/531210 (April 5, 2020).

82. E. Nagy, L. E. Maquat, A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.* **23**, 198–199 (1998).

83. F.-O. Desmet, *et al.*, Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37**, e67 (2009).

84. R. G. H. Lindeboom, F. Supek, B. Lehner, The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).

85. D. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

86. K. Kin, M. C. Nnamani, V. J. Lynch, E. Michaelides, G. P. Wagner, Cell-type Phylogenetics and the Origin of Endometrial Stromal Cells. *Cell Rep.* **10**, 1398–1409 (2015).

87. J. Lonsdale, *et al.*, The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

88. W.-H. Chen, G. Lu, X. Chen, X.-M. Zhao, P. Bork, OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res.* **45**, D940–D944 (2017).

89. J. X. Chong, *et al.*, The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97**, 199–215 (2015).

90. , OMIM Download (October 13, 2017).

91. The 1000 Genomes Project Consortium, *et al.*, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

## Legends for figures

**Figure 1. Functional consequences of LoF variants defining the set of dispensable protein-coding genes.**

Barplots show the distribution of LoF variants defining the set of dispensable genes according to their molecular consequences (stop-gains, frameshifts and splice-disrupting variants, SDVs) and the predicted severity of their functional impact: low probability (light gray) and high probably LoF (dark red; **Methods**).

**Figure 2. Distribution of the set of dispensable genes in organ-expressed genes and adaptive and innate leukocytes-expressed genes, defined from gene expression datasets.**

Relative enrichment in dispensable genes among different gene sets defined from expression datasets based on RNA-seq expression data from the Illumina Body Map project (IBM), the GTEX project and the BluePrint project (adaptive and innate leukocytes). Organ-expressed genes are further classified into two subcategories: organ-specific and organ-pervasive genes (see text). The ratio of dispensable genes versus the total size of each category is indicated. Results are presented separately for 17948 non-OR genes (**A**) and 355 OR genes (**B**) for which expression data could be retrieved, from an initial list of 20232 protein-coding genes (**Methods**). *P*-values for two-tailed Fisher's exact tests comparing the fraction of dispensable genes among the gene subsets against the reference background are reported in the text.

**Figure 3. Distribution of dispensable genes across the five human populations considered**

The numbers of dispensable non-OR genes (**A**) and dispensable OR genes (**B**) in each population (>1% homozygous LoF frequency in a given population) are represented across five categories, indicating whether the gene is dispensable in all five populations considered, four, three or two of them, or is a population-specific dispensable gene. The five populations considered were: Africans (including African American, AFR), Americans (AMR), East Asians (EAS), Europeans (Finnish and Non-Finnish; EUR) and South Asians (SAS). The homozygous LoF variant frequencies were taken from the GnomAD dataset for the purposes of this analysis.

**Figure 4. Distribution of functional scores relating to gene essentiality/redundancy in the *stringent* set of dispensable genes.**

Score distributions presented include: **A**. Gene damage index (GDI). **B**. Residual variation intolerance scores (RVIS). **C.** Probability of being intolerant to both heterozygous and homozygous LoF variants (pLI). **D.** Probability of being intolerant to homozygous LoF variants (pRec). **E**. Proportion of non-lethal non-synonymous mutations, *f*, estimated by SnIPRE. **F.** Inter-species conservation, estimated by GerpRS (**Methods**). Panels display the distribution of scores across dispensable non-OR genes (light green), non-dispensable non-OR genes (dark green), dispensable OR genes (light purple) and non-dispensable OR genes (dark purple). Two-tailed Wilcoxon test *p*-values comparing the distribution of dispensable non-OR genes to that of non-dispensable non-OR genes: (A) *p*-value=6.52e-35; (B) *p*-value=1.30e-37;(C) *p*-value=1.55e-22; (D) *p*-value=2.94e-01; (E) *p*-value=1.51e-29; (F) *p*-value=9.70e-28. Two-tailed Wilcoxon test *p*-values comparing the distribution of dispensable OR genes

to that of non-dispensable OR genes: (A) *p*-value=3.04e-06; (B) *p*-value=1.96e-02; (C) *p*-value=4.64e-01; (D) *p*-value=7.98e-01; (E) *p*-value=6.81e-02; (F) *p*-value=1.85e-02.

**Figure 5.** Evidence for positive selection on common LoF alleles. **A.** Empirical *p*-values for 32 LoF mutations presenting $F_{ST}$ scores measuring allele frequency differentiation, in the 95$^{th}$ percentile of highest values genome-wide, in at least one population. **B.** Empirical *p*-values for 15 LoF mutations presenting integrated haplotype scores (iHS) below the 5$^{th}$ percentile of the lowest values genome-wide (i.e. selection on the LoF allele), in at least one population. $F_{ST}$ and iHS values are reported for each of the 26 populations of the 1000 Genomes Project, grouped into five super populations: AFR (brown), AMR (orange), EAS (purple), EUR (blue) and SAS (pink). Eight common LoF variants with both high $F_{ST}$ and low iHS are highlighted. The color gradients indicate the significance of the *p*-values and only polymorphic sites involving common biallelic LoF SNPs from the stringent set predicted to have severe damaging functional consequences were considered.